基于圈结构的 LPANNI 优化算法 *

刘　继[a,b]，贾芳弟[a]

(新疆财经大学 a.统计与数据科学学院; b.新疆社会经济统计与大数据应用研究中心，乌鲁木齐 830012)

摘　要：针对重叠社区发现准确率提升问题，提出了一种基于圈结构的 LPANNI 优化算法 CLPANNI(Cycle Label Propagation Algorithm with Neighbor Node Influence)。该算法通过挖掘节点的最小圈信息，依据圈比指标衡量节点的重要性并按升序进行标签更新，增加了标签传播过程的稳定性，按照邻居节点影响力大小加权接收邻居节点传递的标签。与 4 种基准算法在 NMI_LFK，NMI_MGH，MOV 指标下进行测试比较，CLPANNI 算法在社区发现准确率方面表现较好。实验结果表明该算法能够有效探测网络重叠社团结构，发现网络的紧密子团，识别的社团分布与真实网络结构更为接近。

关键词：复杂网络；圈结构；标签传播算法；重叠社区发现

中图分类号：TP301.6　　doi: 10.19734/j.issn.1001-3695.2022.02.0078

LPANNI optimization algorithm based on cycle structure

Liu Ji[a,b], Jia Fangdi[a]

(a. *School of Statistics & Data Science*, *b. Xinjiang Social & Economic Statistics & Big Data Application Research Center*, *Xinjiang University of Finance & economics*, *Urumqi* 830012, *China*)

Abstract: In order to improve the accuracy of overlapping community detection, this paper proposed a LPANNI optimization algorithm CLPANNI (Cycle Label Propagation Algorithm with Neighbor Node influence) based on cycle structure by mining the minimum circle information of nodes, measuring the importance of nodes according to the circle ratio index and updating labels in ascending order, the algorithm increases the stability of label propagation process, and receives the labels transmitted by neighbor nodes according to the influence of neighbor nodes. With four benchmark algorithms in NMI_ LFK, NMI_ Through the test and comparison under MGH and Mov indicators, CLPANNI algorithm performs well in the accuracy of community discovery. Experimental results show that the algorithm can effectively detect the overlapping community structure of the network, find the close sub clusters of the network, and the identified community distribution is closer to the real network structure.

Key words: complex network; cycle structure; label propagation algorithm; overlapping community detection

0　引言

复杂网络是人们用来理解现实世界复杂系统的一种抽象模型。它将复杂系统中的实体抽象成节点，将实体之间的关系抽象成连线。社团结构(Community Structure)是复杂网络中最普遍也是最重要的拓扑特性之一，表现为社区内部结构紧密，社区之间连接稀松。社区发现(community detection)对探索复杂系统的运行机制及其功能特性具有重要意义，从是否考虑节点的多社区归属性这一角度，可以将其分为两类，非重叠社区发现算法(Non-overlapping community detection)和重叠社区发现算法(Overlapping community detection)。真实网络中社区结构之间普遍具有重叠区域，往往存在重叠节点，即一个节点同时存在于两个以上社团的现象[1]。而重叠节点对网络结构的演变起到十分重要的促进作用，在万物互联的大数据时代，重叠节点在网络动力学演化中的作用值得深入分析。

自 2005 年以来，学者们试图从不同角度[2~6]设计重叠社区发现探测算法提高重叠社团的识别率和计算效率。其中，标签传播算法[7]以线性时间复杂度优势被各国学者广泛应用到真实网络重叠社区结构探测研究中[8]，COPRA(Community Overlap PRopagation Algorithm)[9]算法是第一个应用于重叠社区发现的标签传播算法，该算法依据节点在不同社团隶属系数的变化识别社团。随后学者们分别提出 SLPA(Speaker-Listener Label Propagation Algorithm)[10]，DEMON(Democratic Estimate of the Modular Organization of a Network)[11]，ACSLPA(Active Semi-supervised SLPA)[12]等算法识别重叠社团，Vinícius da Fonseca Vieira 等人[13]对 5 种表现较好的经典算法 CPM(Clique Percorlation Method)[1]，COPRA[9]，DEMON[11]，SLPA[10]，BigCLAM(CLuster Affiliation Model for Big networks)[14]进行了结构识别效果对比，发现算法识别的社团只是算法运行的结果，并不代表真实的社团，他们指出仅依据常见指标评价算法的优劣存在问题，建议在设计重叠社区发现算法时更多关注重叠区域的节点数量和节点的隶属度等信息。

值得注意的是，以往大部分研究主要基于节点的邻接关系研究网络的功能与特性，但在实际交互场景中往往存在多个节点的复杂相互作用。不论是在自然界还是在虚拟的网络社交圈中，单个个体的行为往往和群体有一定关联，为了更好地协同整体，个体不仅需要考虑个体间的相互作用关系，还需要注意与群体的相互作用关系。考虑到反馈机制对现实网络动态演变的影响，尤其是重要的重叠节点在不同社团的正负反馈作用，本文需要新的视角来分析节点的影响力。

收稿日期：2022-02-28；**修回日期：**2022-04-21　**基金项目：**国家自然科学基金项目(72164034，71762028)；新疆维吾尔自治区高校科研计划项目(XJEDU2019SI006)

作者简介：刘继(1974-)，男，四川达州人，教授，硕导，博士，主要研究方向为数据智能分析、网络社区发现(liuji5000@126.com)；贾芳弟(1995-)，女，甘肃天水人，硕士研究生，主要研究方向为数据智能分析、网络社区发现.

1　相关知识介绍

1.1　重叠社区发现

重叠节点是影响网络拓扑结构演变的关键要素之一，对网络动力学与网络结构相互关系的探索十分重要。因此重叠社团检测算法研究是研究网络动力学有效的方法。本文关注的重叠社团探测算法主要基于团渗流思想和标签传播思想两类，节点依据一定的传播规则自底向上逐步集聚成团，此类算法不同于模块度优化等其他自顶向下的算法，对网络社团结构的质量不做特别限制，因此更符合真实世界中自动生成的组织或者团簇。当前，网络科学迈入新的研究阶段，高阶相互作用动力学将引起人们的极大兴趣，重叠社团探测迎来了挑战，在探测社团结构时需要考虑多节点间的交互特征与网络中观尺度的邻域信息。

1.2　相关工作

文献[15,16]借鉴庞加莱的”剖分”思想，将网络分解为全齐性子网络，开创了网络科学研究的新框架：圈结构。由于圈结构建立了个体与群体的联系，在一定程度上考虑了多个节点的相互作用，可以反映节点的局域影响力，能够进一步指导网络社区发现。文献[17]基于网络的一阶圈结构设计了新的节点重要性指标：圈比，基于圈比指标找到的重要节点分布较为分散，这些重要节点传播高效，不冗余，同步能力强。文献[18]提出了 LPANNI 重叠社区发现算法，该算法融合了 COPRA 和 DLPA(Dominant Label Propagation Algorithm)[19]算法的优点，巧妙的解决了 COPRA 算法在不同网络中参数难以确定的问题；同时充分考虑节点的局域信息，通过综合节点重要性、邻域节点相似性以及邻居节点影响力的方式降低了标签传播算法的随机性；引进历史标签偏好策略，确定节点每次迭代的主标签，增加了重叠社区识别精度。

1.3　评价指标

在不知网络重叠社团结构时，一般用质量函数衡量社团的紧密度，常见的有 EQ[20]，Q_{ov}[21]，M^{ov}[22]。本文选用 M^{ov} 指标，该指标依据每一个节点在不同社团的归属强度来进行计算该节点对社团的贡献程度，是一种非常精确的重叠度衡量办法，这与文献[13]的建议十分吻合。根据节点在社团内连边与社团外连边数目的差值衡量节点对一社团的贡献度，有效避开了其他重叠社区发现模块度指标对高度重叠社团结构的低分辨问题。具体公式如下：

$$\sum_{r=1}^{K} a_{i,c_r} = 1,\ \ a_{i,c_r} \in [0,1]$$
$$M^{ov} = \frac{1}{n_{c_r}} \sum \frac{\sum\limits_{j \in c_r, i \neq j} a_{ij} - \sum\limits_{j \notin c_r} a_{ij}}{d_i \bullet s_i} \bullet \frac{n^e_{c_r}}{\binom{n_{c_r}}{2}}, M^{ov} \in [-1.1] \tag{1}$$

其中：n_{c_r}、$n^e_{c_r}$ 分别代表第 r 个社团 c 的节点数和连边数，由于第一个因子的取值范围在-1 和 1 之间，第二个因子的取值范围在 0 到 1 之间，因此 M^{ov} 的取值在-1 和 1 之间变化。

在已知真实社团结构时，常用 NMI 指标进行社团划分结果的衡量，本文选用 CDlib 库(Community Discovery Library)[23]中的两种 NMI 指标。

2　LPANNI 算法框架

LPANNI 算法首先以固定顺序更新节点的标签，有效解决了标签震荡问题，降低了随机性；其次，合理运用节点的局部信息测度了不同邻居节点的影响力大小；再次，在标签传播过程中只传播社区归属系数最大的主标签，若具有多个相同的最大的主标签则选择历史迭代中出现的主标签，过滤了不重要的标签信息；在算法收敛后，根据节点的标签集信息确定重叠节点。

2.1　符号说明

本文涉及到的关键符号及其含义如表 1 所示。

表 1　符号说明

Tab. 1　Symbol description

符号	含义	符号	含义
CR	圈比	Ld	节点 i 的主标签
Sim	相似度	p	路径长度
NNI	邻居节点影响力	a	路径长度阈值
hl	历史标签偏好	T	最大迭代次数
VQ	更新顺序	Ng(i)	节点 i 的邻居
LNg	节点的主标签集	| L'|	标签集大小
On	重叠节点数	Om	重叠节点社团数
b(c,i)	节点 i 在社团 C 中的隶属系数	l(Cv,bv)	邻居节点 v 的主标签及对应社团的隶属系数

上述符号主要基于本文提出的 CLPANNI 算法，部分符号为后续实验中出现的参数。

2.2　参数初始化

最大迭代次数 T；节点数量 V；迭代时刻 t；用一组有序数对表示节点 i 在不同社区中的隶属强度：b_c(c，i)；节点 i 的邻居节点 NG(i)；节点 i 标签集的隶属系数最大的标签为主标签 D_i，节点 i 的标签集大小|L'|；节点 i 的标签集合 L_i。

初始时刻，网络 G(V,E)中的节点各自为一个独立社区，节点的隶属系数为 1，即社区 i 的隶属系数 b_i 为 1，记作隶属 b_i(i，1)。

2.3　更新策略

```
输入: G=(V,E,w),最大迭代次数 T。
输出:社区识别结果。
阶段 1: 固定标签更新顺序
for i in V:
    依据节点重要性公式计算节点的重要性
    依据节点相似性公式衡量节点间的相似性
    依据邻居节点影响力公式计算邻居节点的重要性
end for
按节点重要性的大小或者序列号的大小升序排列为 VQ
阶段 2: 标签传播过程
t=0
for i in V:
    l[i] = {i,1};
    主标签 Dᵢ = i;
end for
while t < T:
    for i in VQ:
        LNg = {l(c₁,b₁),l(c₂,b₂),...,l(cᵥ,bᵥ)},v∈Ng(i);
        L' = 按照更新规则更新节点 i 的标签集
        for ls in L':
            if b' < 1/|L'|:
                then delete ls from L';
            end if
        end for
        Lᵢ= 归一化的节点标签 L'
        确定节点 i 本次迭代后的主标签 Dᵢ;
    end for
    若所有节点的标签集大小以及主标签不再变化
        则停止迭代;
    end if
    t = t+1;
end while
Output Lᵢ of each node i,(i∈V)
```

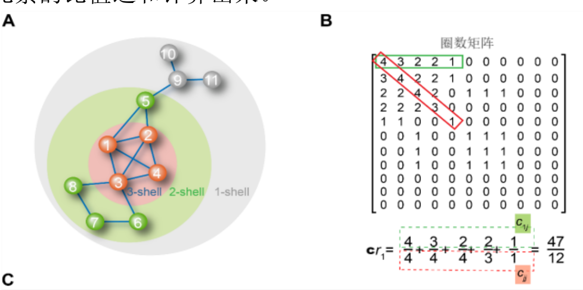
LPANNI 算法包含 2 个阶段的处理，即固定标签更新顺序和标签传播过程。阶段 1 主要按照邻居节点的影响力大小更新标签信息，阶段 2 中在确定主标签时，如果只有一个最大的隶属度，则传播该社区标签；如果存在多个相同的最大隶属度社区标签，则优先选择上次迭代中的主标签，否则随机选择一个作为主标签。

LPANNI 算法考虑了节点的局域信息，通过计算邻居节点的影响力巧妙设计了标签更新规则，在传播规则上具有很大的借鉴意义。但只关注节点对的相互作用，主要借助节点的度信息设计相关公式，对具有相同结构的节点区分度不大，需要借助节点 ID 顺序较多，本文基于以上优点和不足对其进行改进。

3　CLPANNI 算法设计

事物有从简单到复杂的一个发展过程，错综复杂的交互关系使得组织得以延续壮大，组织间的交互带来了联通与演化。以网络科学视角可以看到一些特定的网络结构，例如星结构，链结构，以及圈结构。圈结构是构成网络的基本结构之一，是形成网络功能的最重要机制之一：反馈效应的结构基础，而反馈对事物的发展演化十分关键。

在网络动力学同步的研究中，史定华等人发现最容易同步的网络是度数相同、路和相同并且最小、周长相同并且最大的几乎全齐性网络[15,16]。圈在结构上给网络连通带来了冗余路径，在功能上表征了反馈机制，在网络动力学中产生了强化效应，很容易增强社会协同效应，因此圈结构在保持网络连通性和维护网络的动态交互方面比较重要。在此基础上，范天龙等人认为参与许多圈的节点很重要，这些节点对网络的连通、同步以及控制方面有极大的影响；他们基于网络最小圈设计了基于圈结构的节点重要性指标：圈比[17]。具体来看：图 1 的子图 b 中计算了子图 a 中节点 1 的圈比，子图 c 是所有节点的度、H 指数、核数、圈比值、中介中心性等信息。他们对网络的一阶圈结构的最小圈定义了一个新的矩阵，称为圈数矩阵。圈数矩阵的阶数与网络中的节点数相同，矩阵的第 i 行(列)描述了节点 i 与其他节点的共圈情况，矩阵的元素表示网络中任意两个节点之间的共圈数量。这样，节点 i 的圈比值就可以根据圈数矩阵中第 i 行非零元素与对角线元素的比值之和计算出来。



Each node's associated cycles in *S* and the four indices under consideration.

Node label	Associated cycles in *S*	Degree	H-index	Coreness	Cycle ratio	Betweenness	Articulation ranking
1	{1, 2, 3}, {1, 2, 4}, {1, 4, 3}, {1, 5, 2}	4	3	3	47/12	0.22	0
2	{2, 3, 1}, {2, 4, 1}, {2, 1, 5}, {2, 4, 3}	4	3	3	47/12	0.22	0
3	{3, 1, 2}, {3, 1, 4}, {3, 2, 4}, {3, 6, 7, 8}	5	3	3	68/12	0.48	1
4	{4, 1, 2}, {4, 1, 3}, {4, 2, 3}	3	3	3	10/4	0	0
5	{5, 2, 1}	3	3	2	6/4	0.47	1
6	{6, 7, 8, 3}	2	2	2	13/4	0.09	0
7	{7, 8, 3, 6}	2	2	2	13/4	0.01	0
8	{8, 3, 6, 7}	2	2	2	13/4	0.09	0
9	Null	3	1	1	0	0.38	2
10	Null	1	1	1	0	0	0
11	Null	1	1	1	0	0	0

图 1　圈比计算(图片修改自文献[17])

Fig. 1　Calculation of cycle ratio (picture from literature[17])

圈提供的冗余连通和反馈机制，使得圈上节点无论在同步还是传播中[24]，都有更高的概率被接触和同步，能更好模拟社会增强效应。考虑到圈比指标在挖掘高传播节点方面的突出表现，本文将圈比运用到标签传播算法中来，先借助圈比找到网络中传播能力强的节点，并运用 LPANNI[18]算法的标签传播策略来提高网络社团的划分精度。

3.1　相关定义

简单无向网络 G(V，E)，其中 V 和 E 分别表示节点集和连边集，n 表示节点的数量，m 表示连边的数量。

圈(Cycle)：二维平面上具有相同起点和终点的闭合路径，圈的大小等于它的连线数目，最小圈是指含该节点的最小环路。

周长(Girth)：从该点出发再返回它的最短环路所含的连线数目，即经过该节点的最小圈长。

圈数(Cycle number)：含有该节点的最小圈的数量。

圈比(Cycle Ratio：CR)：节点重要性衡量指标，圈数矩阵中第 i 个节点所在行的元素比上对应行的对角元素之和，得到 i 节点的圈比，具体计算公式为

$$\mathrm{CR}_i=\begin{cases}0.1, c_{ii}=0\\ \sum_{j,c_{ij}>0}\frac{c_{ij}}{c_{jj}}, c_{ii}>0\end{cases}\tag{2}$$

C_{ij} 是 S 中节点 i 和 j(i≠j)的共圈数量。若 i=j，则 c_{ii} 是 S 中包含节点 i 的圈数，CR_i 为 i 节点的圈比值，为了能够精确衡量邻居节点影响力的大小，故这里将第一种情况取为 0.1。

节点相似性(Similarity：Sim)：本文衡量节点相似性主要基于网络结构，文献[25]对局部相似性指标的相关研究进行了梳理，并分析了这些指标的设计原理。指出结构相似性指标可以分为基于局部信息、路径及随机游走三类，文献[18]提出的相似性指标结合了节点的局部信息以及路径长度，有效融合了二者优势，故本文沿用该相似性指标，具体计算公式为

$$\mathrm{Sim}(x,y)=\frac{s(x,y)}{\sqrt{\sum_{u\in N_{g_x}}s(x,u)*\sum_{v\in N_{g_y}}s(y,v)}}\tag{3}$$

其中，$s(i,j)=\sum_{|p|=1}^{a}\frac{(A^{|p|})_{ij}}{|p|}$ 。

上式中，p 表示直接或间接连接节点 i 和节点 j 的路径。|p|表示 p 的长度，它在 1 到 a 之间变化。$|A|^{p}|$表示 p 的测度度量。路径长度阈值 a 来控制计算复杂度，用来区分两个节点因度值差异对节点相似度带来的影响。

路径长度阈值 a 来控制计算复杂度，用来区分两个节点因度值差异对节点相似度带来的影响。邻居节点影响力(Neighbor Node Influence：NNI)：考虑到邻居节点由于具有不同的局部结构，对节点的影响力也不尽相同，在标签传递时需要测度不同邻居节点的差异。文献[18]提出的 NNI 综合考虑邻居节点的重要性大小和邻居节点与该节点的相似性程度，相对来说比较客观，本文沿用。具体公式为

$$NNI_y(x)=\sqrt{CR(y)*\frac{Sim(x,y)}{\max_{h\in Ng(x)}Sim(h,x)}}\tag{4}$$

3.2　LPANNI 算法的改进

LPANNI 算法仅仅通过节点的度和三角形信息设计节点重要性公式来衡量节点的重要性以及邻居节点的影响力大小，没有考虑更多的圈结构，因此看到的信息是十分有限的，不足以衡量具有相同局部结构节点的差异。而圈比指标通过衡量节点参与邻居节点圈的程度识别重要节点，有利于标签的动力学传播，本文基于圈结构信息对无向无权网络提出重叠社区发现算法 CLPANNI，根据圈比的升序固定节点的标签更新顺序，提高社团识别的精度。

3.3　CLPANNI 算法框架

CLPANNI 算法主要分为 2 个阶段，第一个阶段完成节点圈比和邻居节点影响力的计算，第二个阶段进行标签传播，找到全部节点的隶属社团，输出节点的标签集。具体步骤如图 2 所示。左边一部分为 CLPANNI 算法的第 1 阶段，右边一部分为 CLPANNI 算法的第 2 阶段。

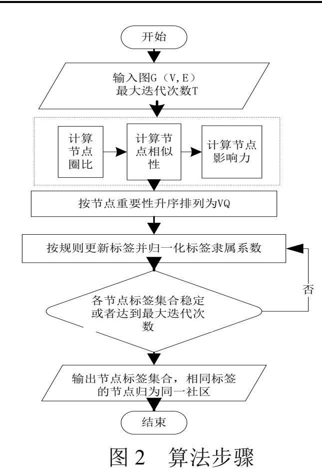


图 2　算法步骤

Fig. 2　Algorithm steps

3.4　标签传播规则

文献[9]提出的 COPRA 算法考虑节点在不同社团的隶属强度，节点在不同社团的隶属强度类似于一个个体在不同层面中注意力或者精力的分散程度，总和加起来为 1。文献[18]以节点重要性的升序固定标签传播序列增加了算法的稳定性，通过邻居节点影响的标签更新策略和历史标签偏好策略提出的 LPANNI 算法解决了 COPRA 算法需要提前设置节点最多拥有 v 个标签的缺陷，这里本文以节点的圈视角衡量节点的重要性程度，具体标签传播规则为

阶段 1：根据节点的圈比值，从小到大对节点排序，若节点的圈比值相同，则以 ID 大小升序排列，得到固定的更新序列 VQ。

阶段 2：依据 VQ 的顺序进行标签更新。初始时刻，每个节点都完全属于自己，即有 $Ld_i=(i,1)$。接下来则按 VQ 顺序接收邻居节点的主标签后形成 LNg 标签集，主标签是指邻居节点指传递最大的隶属系数及社团标签。

$$LNg=\{l(c_1,b_1),l(c_2,b_2),...l(c_v,bv)\},v\in Ng(i) \tag{5}$$

并按照邻居节点的影响力大小 NNI 对隶属系数进行加权处理，得到大小为邻居节点数的新标签集 L'，此时每个节点的总隶属系数为 1。加权处理传递的社团隶属系数

$$b''(c,i)=\frac{b'(c,i)}{\sum_{l(c,b')\in L''}b'(c,i)},\ \sum_{l(c,b'')\in L''}b''(c,i)=1 \tag{6}$$

加权处理后的标签集

$$L'=\{l(c_1,b_1'),l(c_2,b_2'),\cdots,l(c_{|L'|},b_{|L'|}'),\},\ \sum_{l(c,b')\in L'}b'(c,i)=1 \tag{7}$$

自适应删除无用标签 $b'(c,i)<1/|L'|$，归一化后得到此次迭代的标签集 L”。如此迭代 T 次后输出各节点的标签集。

$$b'(c,i)=\frac{\sum_{l(c_v,b_v)\in LNg(i),v\in Ng(i),c_v=c}b(c_v,v)*NNI_v(i)}{\sum_{l(c_v,b_v)\in LNg(i),v\in Ng(i)}b(c_v,v)*NNI_v(i)} \tag{8}$$

识别具有最大隶属系数的标签为节点 i 的主标签，若有多个主标签，则选择上一步迭代的主标签，否则随机选择一个作为主标签。当全部节点的标签集和主标签稳定时，停止迭代，输出节点的标签集。

4　实验

4.1　实验数据

1)人工数据集

LFR 人工数据集能够合成接近真实情况的使得节点数和社团数均满足幂律分布的网络。因此本文使用 LFR 基准网络生成数据，分别用清晰度 μ 为 0.1 或 0.3 的两组数据进行实验对照。每一组数据中分 3 个等级，节点数目分别为 1000，3000，5000，每一个等级中分了 5 组不同重叠程度的数据。其具体信息如表 2 所示。

2) 合成网络的最小圈

以往重叠社团发现算法很少关注网络的圈结构分布情况，部分算法[18,26,27]考虑网络的三角形信息。对 LFR 数据进行圈结构分布的可视化有助于理解进一步了解网络中的圈结构信息。在运用 CLPANNI 算法检测网络的社团结构的同时，可以清楚的看到网络的最小圈分布情况，如图 3 所示。

表 2　合成网络的具体参数

Tab. 2　Specific parameters of synthetic network

数据集	节点数量	社团规模	μ	On	Om
LFR1	1000	10-80	0.1	30	1~8
LFR2	3000	10-80	0.1	90	1~8
LFR3	5000	10-80	0.1	150	1~8
LFR4	1000	10-80	0.3	30	1~8
LFR5	3000	10-80	0.3	90	1~8
LFR6	5000	10-80	0.3	150	1~8

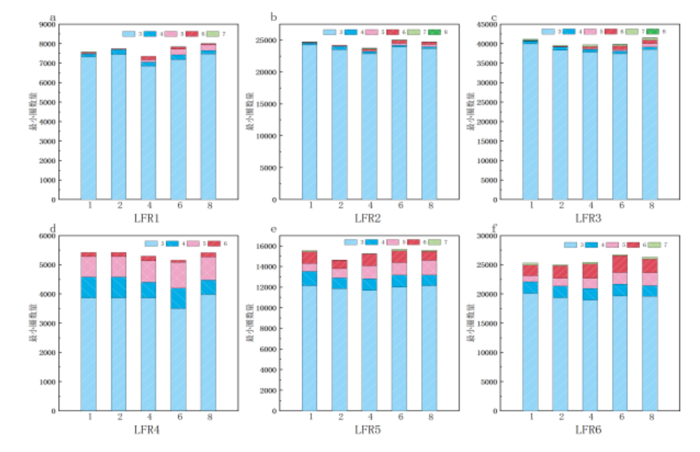


图 3　合成网络的最小圈分布

Fig. 3　Minimum cycle distribution of the synthetic network

由图可知，本实验的 6 个 LFR 网络中含较多的圈结构数据，不同规模的数据具有不同的圈结构分布，最小圈的周长为 3，最大圈的周长为 8，且最小圈中三角形数量占比过半。在混合参数 μ 为 0.1 时，三角形在网络最小圈数量占比至少达到的 90%；当混合参数 μ 为 0.3 时，随网络结构清晰度的下降，最小圈数量明显减少，三角形的占比有所下降，四边形与五边形的占比明显上升，这在一定程度上印证了当网络结构不明显时社团识别效果不好的原因。当网络规模不变时，随网络拓扑结构复杂程度上升最小圈的种类和分布多样化特征更为突出，这就加大了社团识别的难度。

3)真实数据集

文献[12]提供了已知社团信息的 3 个真实数据，分别是处理过的共同购买网络 Amazon，科学家合作网络 DBLP，友谊网络 YouTube，并列出了网络的重叠节点信息。本文对这 3 个真实网络数据进行了最小圈挖掘，发现 Amazon 数据与 DBLP 数据中的最小圈种类较少，但 YouTube 数据中的圈分布具有多样性的特征，其中最小的圈周长为 3~10。说明 YouTube 网络结构特征相对比较复杂。具体相关信息如表 3 所示。

表 3　真实网络的具体信息

Tab. 3　Specific information about the real network

数据集	Amazon	DBLP	YouTube
节点数	7411	7233	6426
边数	21214	33045	23226
社团数	876	613	1058
最大社团规模	27	38	31
最小社团规模	5	10	5
重叠节点最大社团数	4	8	11
最小圈总数	26171	108882	40517
最小圈分布	{(3,26061),(4,110)}	{(3,108880),(6,2)}	{3,4,5,6,7,8,9,10}
重叠节点数(占比)	1394(18%)	214(3.3%)	865(13%)

由于 YouTube 数据中的最小圈数量较大且种类较多，不便枚举。这里将其圈分布展开分析，如图 4 所示。

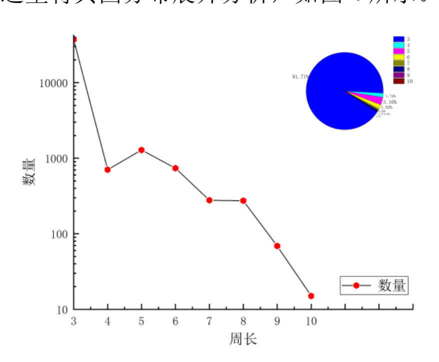


图 4　YouTube 网络的最小圈分布

Fig. 4　Minimum circle distribution of YouTube network

可以看到 YouTube 网络中，周长为 3 的圈：三角形依然是最小圈的主体组成部分，数量级别最高。其次为周长为 4-8 的圈，在该网络中最长的圈周长为 10。

4.2　实验结果

以 SLPA，DEMON，CPM 和 LPANNI 为对比算法，经过多次实验调参，相关算法的参数设置如表 4 所示。

表 4　基准算法的具体参数

Tab. 4　Specific parameters of the benchmark algorithm

数据集	Amazon	DBLP	Youtube
CPM	k=3	k=3	k=3
DEMON	min_com_size=3,epsilon=0.25		
SLPA	T=21,r=0.1		T=21,r=0.2
LPANNI	T=20,a=3,b=0		
CLPANNI	T=20,a=3,b=0		

其中，由于 SLPA 算法不稳定，经过 10 次重复实验取平均得到具体值，CPM 算法中参数 k 一般取 3~6，经过测试发现将 k 取为 3，得到的各项测试结果更好，其他算法中的参数按照 CDlib 库[23]的默认值进行测试。

先对 LFR 数据进行测试，NMI_LFK[28]是学者常用的测试指标，该指标是标准互信息在重叠社区发现中的拓展，但有时候会高估两个社团的相似性。NMI_MGH[29](也称 NMI_{max})是 Aaron F. McDaid 等人对 NMI_LFK 作出的优化指标，本文选用该指标进行测试。

由图 5 可知，CLPANNI 算法与 LPANNI 算法相对其他基准算法表现较好，在网络规模增大，网络社团结构清晰度下降的情况下，还能准确识别网络的社团结构。当重叠节点数量增加时，CLPANNI 算法的表现略强于 LPANNI 算法，说明本文的改进有效，圈比在挖掘节点重要方面表现更好。

为进一步验证 CLPANNI 算法的准确度，对 3 个社团结构已知的真实数据进行测试，如图 6 所示。

在真实数据的实验结果中，可以看到在 3 种指标的检测下，除了 YouTube 数据，CLPANNI 算法较 LPANNI 算法测试结果表现稍低外，在其他数据的表现比 LPANNI 算法好。在 M^{ov} 指标下，CLPANNI 算法表现较 LPANNI 算法好，尤其在 Amazon 数据集中优势更为明显。SLPA 算法在 DBLP 数据上表现相对较好，说明该算法在社团结构明显的网络上表现较好。

另一方面，也说明本文的参数设置合理，但该算法存在极大的不稳定性和随机性，需要多次重复实验，不能保证每次都能得到较好的划分，故并不能提供一个可靠的结果；DEMON 算法在 YouTube 中，NMI_MGH 得分较低，说明民主投票机制的标签传播策略不适用于此类网络；总体来说，CLPANNI 算法表现良好，在多项指标的测试下都有不错的表现。

5　讨论

5.1　算法探测结果

由图 5 的实验结果可知，网络社团结构越模糊，社团间重叠程度越大，CLPANNI 算法相较其他算法表现越好；从图 6 的重叠社团结构评价指标来看，本文所提算法 CLPANNI 在 DBLP 网络与 Amazon 网络表现较原算法更优；另外，多个算法对 YouTube 数据的社团探测效果欠佳，因此有必要对该数据的探测结果进一步分析，探寻其背后的原因。YouTube 数据是社交网络数据，故在该网络上存在一定的社会网络增强效应。(注：由于 SLPA 算法具有随机性，重复实验 15 次。平均来说，能够识别到社团 477 个，这里展示社团数量为中位数的实验结果。)

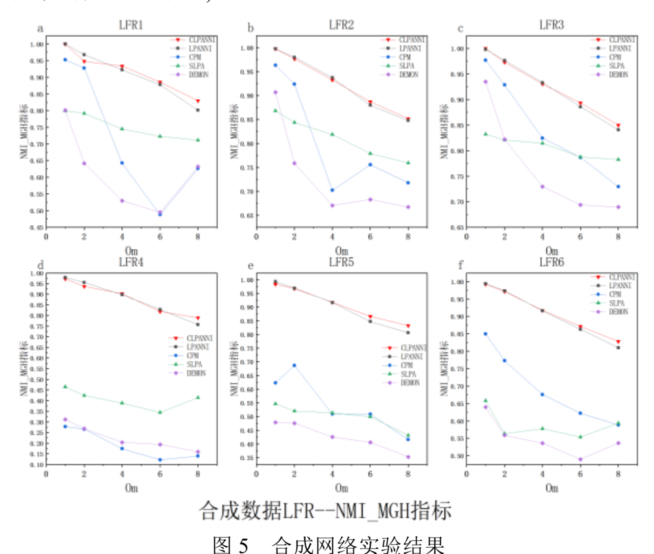


图 5　合成网络实验结果

Fig. 5　Experimental results of synthetic network

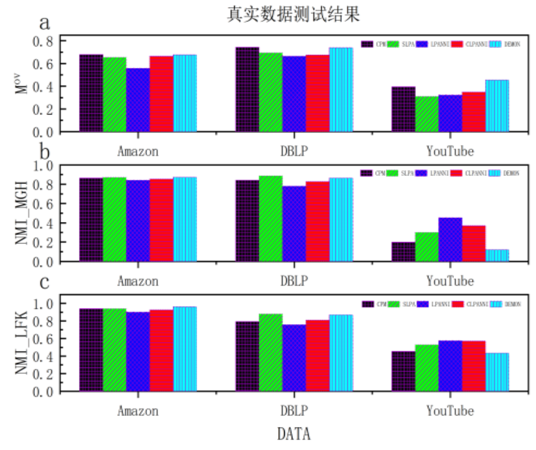


图 6　真实网络实验结果

Fig. 6　Real network experiment results

真实情况下，YouTube 网络 6426 个节点中有重叠节点 865 个，社团 1078 个。主要的社团规模为 5，共有 307 个。表 5 分别展示了各个算法在 YouTube 社交网络上探测的社团重叠模块度、主体社团规模及其探测数量、社团识别数量与总节点频次信息。在本次实验中，SLPA 算法能发现社团 455 个。CPM 算法能够识别到紧密的社团结构，K=3，识别到的结构是以三角形为单元的团簇 230 个。在社团结构识别上，可以看到 DEMON 算法发现的社团数量最少同时 M^{ov} 得分最高，但在重叠节点探测上存在过拟合的问题；相对来说，CPM 与 SLPA 算识别的社团数量较为相近，能够探测出 455 个社团；LPANNI 算法发现的社团数量最多但社团较小；在重叠节点发现方面，CLPANNI 算法的识别效果不如 LPANNI 算法，但 CLPANNI 算法对主体社团规模的识别准确度较高，

说明改进的 CLPANNI 算法较原始算法在社团规模探测方面拟合更好。YouTube 具有多样化的最小圈分布，说明本文算法对连边冗余圈结构复杂的网络探测探测效果更优，较原始算法能更好模拟社会增强效应。

表 5　算法探测结果

Tab. 5　Algorithmic detection results

方法	M^{ov} 指标	主体社团规模及大小	社团探测数量	总节点频次
RealData	0.1890	{5,307}	1058	7775
CPM	0.3959	{3,230}	455	4570
SLPA	0.5145	{5,120}	455	7623
DEMON	0.677	{4,46}	200	13898
LPANNI	0.3233	{2,167}	867	6675
CLPANNI	0.3485	{5,129}	574	6543

5.2　社团规模分布

为进一步分析各个算法在该网络社团结构探测方面的表现，对各个算法探测到的社团数量分布进行对比。

LPANNI,CLPANNI,DEMON,SLPA 算法都含有标签传播算法的思想，因此会存在标签传播算法特定的缺点，存在标签过度传播和大团吃小团的现象，去除奇异值得到图 7 的社团探测结果。CPM 算法主要通过完全子图的渗流识别社团结构，因此得到的社团大小与网络局域结构的紧密性有很大的关联。在真实的 YouTube 数据中，最大的社团含 31 个节点，但 5 种算法识别到的最大社团大小均超过 31，说明该网络存在紧密的联通块。以上实验结果为 CLPANNI 在最宽松的条件下进行社团发现，没有针对性地调参，以上实验结果为 CLPANNI 在最宽松的条件下进行社团发现，没有针对性地调参，如果精心筛选节点隶属度阈值，检测效果还会有提升，但会消耗较多时间。

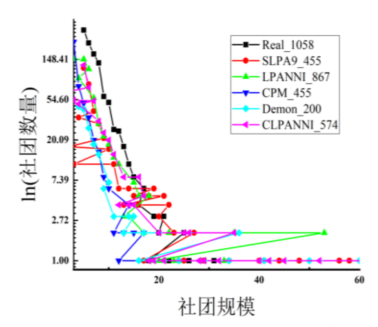


图 7　YouTube 网络社团探测结果

Fig. 7　The results of the youtube community detection

5.3　可视化分析

图 8 为 Gephi 自带的 louvain 算法社团识别结果，颜色相同的为同一社团，带有标签的节点为至少跨 8 个社团的重叠节点。经过可视化发现，YouTube 数据的网络社团结构不清晰，确实存在巨大连通片，且高重叠节点分布比较集中。这或是几个典型算法在该数据集中社团识别效果不佳的部分原因。YouTube 社交网络由于重叠节点分布集中，使得该网络在舆情传播较容易形成规模性的网络意见，在实际应用中应该更关注重叠节点的意见情况，为网络舆情预警与引导提供智力支持。

6　结束语

本文对 LPANNI 算法进行了优化，提出了一种基于网络圈结构信息的检测重叠社团结构的标签传播算法 CLPANNI，该算法提高了具有紧密网络结构的社团识别精度。CLPANNI 不仅具有良好的稳定性，还能在检测重叠社团结构的同时得到网络中每一个节点的最小圈信息以及网络中不同周长的最小圈分布，这有助于进一步了解网络的结构特点，提供中观尺度信息指导重叠社团结构的探测。

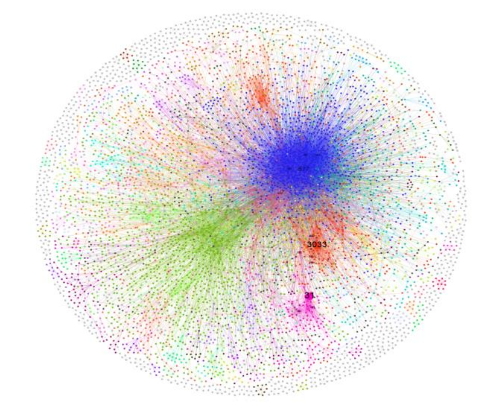


图 8　YouTube 高重叠节点分布

Fig. 8　Distribution of highly overlapping nodes in YouTube

经过对比 YouTube 网络的社团输出信息，本文发现 CLPANNI 算法对真实网络中的重叠节点挖掘、社团结构识别效率方面还有待提升，未来将结合圈结构对具有紧密结构的网络进行深入分析。真实网络往往会随外界环境动态演变，而拓扑结构又影响信息的传输，如何利用探测到的重叠节点及其社团结构分析网络的动态演化特性，值得深入研究。考虑到网络的连接往往是在信息不完全条件下作出的有限选择，未来在研究中还需要融合先验知识，整合网络高阶信息进一步量化网络，合理嵌套节点属性信息挖掘网络的重叠社团结构，发掘关键重叠节点协助预判网络的动态演化方向。

参考文献：

[1] Palla G, Deranyi I, Farkas I, *et al.* Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435 (7043): 814.

[2] 李辉，陈福才，张建朋，吴铮，李邵梅，黄瑞阳. 复杂网络中的社团发现算法综述 [J]. 计算机应用研究, 2021, 38 (06): 1611-1618. DOI: 10. 19734/j. issn. 1001-3695. 2020. 06. 0211. (Li Hui, Chen Fucai, Zhang Jianpeng, *et al.* Survey of community detection algorithms in complex network [J]. Application Research of Computers, 2021, 38 (06): 1611-1618. DOI: 10. 19734/j. issn. 1001-3695. 2020. 06. 0211.)

[3] 许英. 利用改进蚁群算法的重叠社团检测分析方法 [J]. 计算机应用研究, 2020, 37 (05): 1375-1379. DOI: 10. 19734/j. issn. 1001-3695. 2018. 10. 0803. (Xu Ying. Novel algorithm of overlapping community detection and analysis with improved ant colony algorithm [J]. Application Research of Computers, 2020, 37 (05): 1375-1379. DOI: 10. 19734/j. issn. 1001-3695. 2018. 10. 0803.)

[4] 付饶，孟凡荣，邢艳. 基于节点重要性与相似性的重叠社区发现算法 [J]. 计算机工程, 2018, 44 (9): 192-198. (Fu Rao, Meng Fanrong, Xing Yan. Overlapping Community Discovery Algorithm Based on Node Importance and Similarity [J]. Computer Engineering, 2018, 44 (9): 192-198.)

[5] 楚杨杰，杨忠保，洪叶. 局部扩展的遗传优化重叠社区发现方法 [J]. 计算机应用研究, 2019, 36 (04): 1106-1109. (Chu Yangjie, Yang Zhongbao, Hong Ye. Local extension approach through genetic algorithm for overlapping community detection [J]. Application Research of Computers, 2019, 36 (04): 1106-1109.)

[6] Shchur O, S Günnemann. Overlapping Community Detection with Graph Neural Networks [J]. https://arxiv. org/abs/1909. 12201v1

[7] Raghavan U N, Albert R, Kumara S. Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks [J]. Physical Review E, 2007, 76 (3 Pt 2): 036106.

[8] 张应龙，夏学文，徐星，等. 面向标签传播算法的社团检测研究现状及展望 [J]. 小型微型计算机系统, 2021, 42 (5): 1093-1102. (Zhang

Ying-long, Xia Xuewen, Xu Xing, *et al.* Review on label propagation agorithms for community community [J]. Journal of Chinese Computer Systems, 2021, 42 (5): 1093-1102.)

[9] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2009, 12 (10): 2011-2024.

[10] Xie J, Szymanski B K, Liu X. SLPA: Uncovering Overlapping Communities in Social Networks via A Speaker-listener Interaction Dynamic Process [J]. IEEE, 2012.

[11] Coscia M, Rossetti G, Giannotti F, *et al.* DEMON: a Local-First Discovery Method for Overlapping Communities [C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[12] Alghamdi E, Greene D. Active semi-supervised overlapping community finding with pairwise constraints [J]. Applied Network Science, 2019, 4 (1): 1-27

[13] Vieira V F, Xavier C R, Evsukoff A G. A comparative study of overlapping community detection methods from the perspective of the structural properties [J]. Applied Network Science, 2020, 5 (1): 1-42.

[14] Yang J, Leskovec J. Structure and overlaps of ground-truth communities in networks [J]. ACM Transactions on Intelligent Systems and Technology (TIST) , 2014, 5 (2): 1-35.

[15] Shi, D, Chen, *et al.* Searching for Optimal Network Topology with Best Possible Synchronizability [J]. Circuits and Systems Magazine, IEEE, 2013, 13 (1): 66-75.

[16] Shi D, Linyuan L, Chen G. Totally Homogeneous Networks [J]. arXiv: 1903. 11289

[17] Fan, T. , Lü, L. , Shi, D. *et al.* Characterizing cycle structure in complex networks. Commun Phys 4, 272 (2021) . https://doi. org/10. 1038/s42005-021-00781-3

[18] Lu M, Zhang Z, Qu Z, *et al.* LPANNI: Overlapping Community Detection Using Label Propagation in Large-Scale Complex Networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31 (9): 1736-1749.

[19] He-Li, Sun, Jian-Bin, *et al.* Detecting overlapping communities in networks via dominant label propagation [J]. Chinese Physics B, 2015, 24 (1): 24 018703.

[20] Shen H W. Detecting the Overlapping and Hierarchical Community Structure in Networks [J]. Springer Berlin Heidelberg, 2013

[21] Nicosia V, Mangioni G, Carchiolo V, *et al.* Extending the definition of modularity to directed graphs with overlapping communities [J]. Journal of Statistical Mechanics Theory & Experiment, 2009, 2009 (03): 3166-3168.

[22] Lázár A. , Ábel D. , Vicsek T. Modularity measure of networks with overlapping communities [J]. EPL (Europhysics Letters) , 2010, 90 (1): 983-995.

[23] Rossetti G, Milli L, Cazabet R. CDLIB: a python library to extract, compare and evaluate communities from complex networks [J]. Applied Network Science, 2019, 4 (1) .

[24] 范天龙. 网络中的圈结构 [EB/OL]. https://swarma. org/?p=31541. 2021-8-10/2022-4-4. (Fan Tianlong. Cycle structure in a network [EB/OL]. https://swarma. org/?p=31541. 2021-8-10/2022-4-4)

[25] 李艳丽, 周涛. 链路预测中的局部相似性指标 [J]. 电子科技大学学报, 2021, 50 (03): 422-427. (Li Yan-li, Zhou Tao. Local Similarity Indices in Link Prediction [J]. Journal of University of Electronic Science and Technology of China, 2021, 50 (03): 422-427.)

[26] 郑文萍, 毕欣琦, 杨贵. 一种基于非对称三角形割的重叠社区发现算法 [J]. 南京师范大学学报（工程技术版）, 2022, 22 (01): 1-8. (Zheng Wenping, Bi Xinqi, Yang Gui. An overlapping community detection algorithm based on asymmetric triangle cuts [J]. Journal of Nanjing Normal University (Engineering and Technology Edition, 2022, 22 (01): 1-8.)

[27] 潘剑飞, 董一鸿, 陈华辉, 钱江波, 戴明洋. 基于结构紧密性的重叠社区发现算法 [J]. 电子学报, 2019, 47 (01): 145-152. (Pan Jian-fei, Dong Yi-hong, Chen Hua-hui, *et al.* The Overlapping Community Discovery Algorithm Based on Compact Structure [J]. Acta Electronica Sinica, 2019, 47 (01): 145-152.)

[28] Lancichinetti A, Fortunato S, J Kertész. Detecting the overlapping and hierarchical community structure of complex networks [J]. New Journal of Physics, 2009, 11: 033015

[29] Mcdaid A F, Greene D, Hurley N. Normalized Mutual Information to evaluate overlapping community finding algorithms [J]. arXiv: 1110. 2515 (2011) .